aNOF：一种极致低时延方案

华为

刘秦飞

lucas.liuqinfei@huawei.com

目录

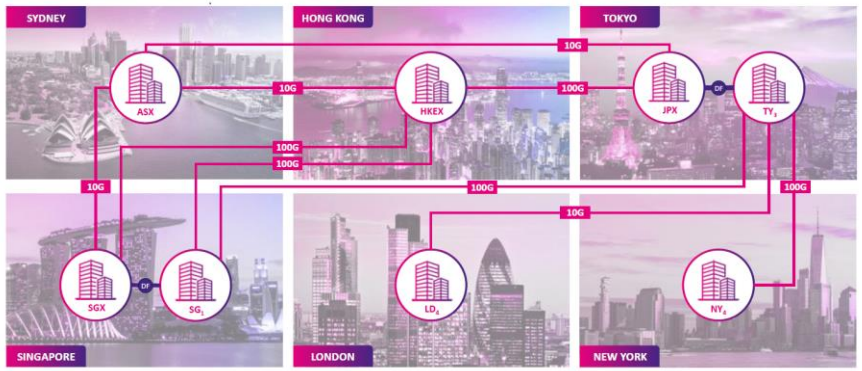
1 技术背景 —— 低时延在存储业务竞争力中的重要性

Best Low Latency Data Feed

- IPC – Connexus ALPHA
- Iress – QuantFEED
- Pico – RedlineFeed
- Quincy Data / McKay Brothers – Quincy Extreme Data (QED) service
- Refinitiv

Any **delay** between receiving market data and making a trade could potentially result in **millions of dollars** in lost revenue and profit.

Enabling Ultra Low Latency Trading in Asia-Pacific



- OLTP领域，尤其在证券高频交易市场（HFT），对于数据写入有着极致低时延（ULL）的诉求。数据中心的数据持久化过程，要求在数据写入主节点后，再将副本写入远端节点（故障域）。因此，单点写入时延影响了数据写入的整体时延。
- 目前，业界已经有rpma等方案实现对远端持久内存的读写，以达到极致低时延的数据传输方案。

主流云盘时延

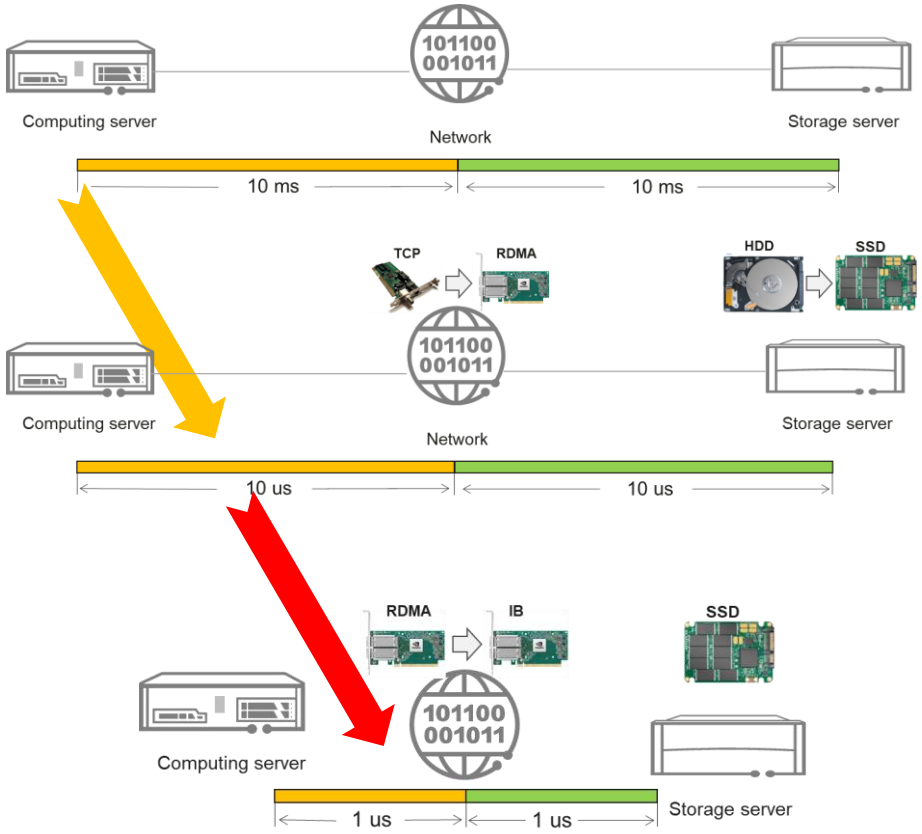
参数	极速型SSD V2（公测）	极速型SSD	超高IO	通用型SSD V2	通用型SSD	高IO	普通IO（上一代产品）
单队列访问时延[d]（ms）	亚毫秒级	亚毫秒级	1	1	1	1~3	5~10

性能类别	ESSD AutoPL云盘	ESSD PL-X云盘	ESSD云盘				SSD云盘	高效云盘	普通云盘
			PL3	PL2	PL1	PL0			
单路随机写平均时延（ms），Block Size=4K	0.2	0.03	0.2	0.2	0.2	0.3~0.5	0.5~2	1~3	5~10

1 技术背景 —— 使用NOF协议实现低时延存储方案

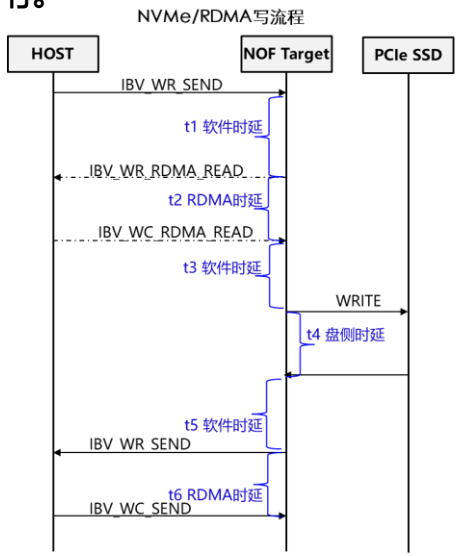
OpenEuler

分布式存储时延 = 网络时延 + 介质时延



NVMe SSD太慢 VS Optane SSD太贵
是否有更实惠高效的方案?

NOF协议是网络存储广泛使用的一种低时延协议，实现了对远端NVMe设备的高性能访问，由于其客户端的CPU开销较低，生态支持比较好，在较多的存储产品中均有应用。



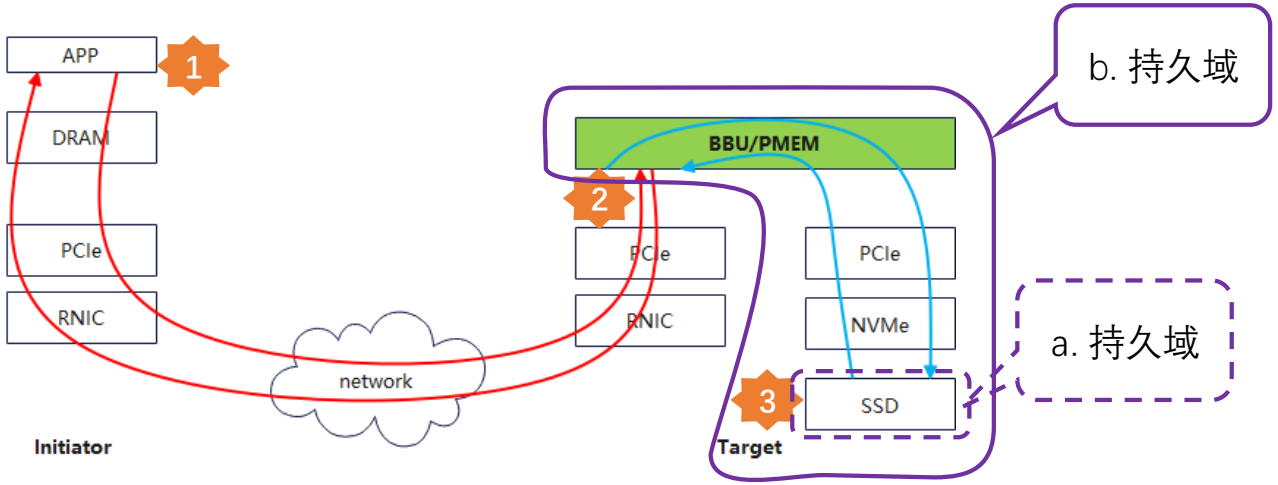
在NOF协议中，时延敏感的小块写一般采用in-capsule的方式，在提交到NVMe控制器后需poll到完成信息，才返回client写成功，此处包含了相对较长的落盘时延。

2 aNOF方案 —— 通过PMEM优化NOF协议

aNOF（Accelerated NVMe-oF）方案，是在NOF协议基础之上，使用硬件应答替换软件应答，极大缩短关键时延路径，从而实现了远端写入数据的极致低时延，并借助PMEM设备来保障数据在异常掉电场景下的可靠性。

关键技术：

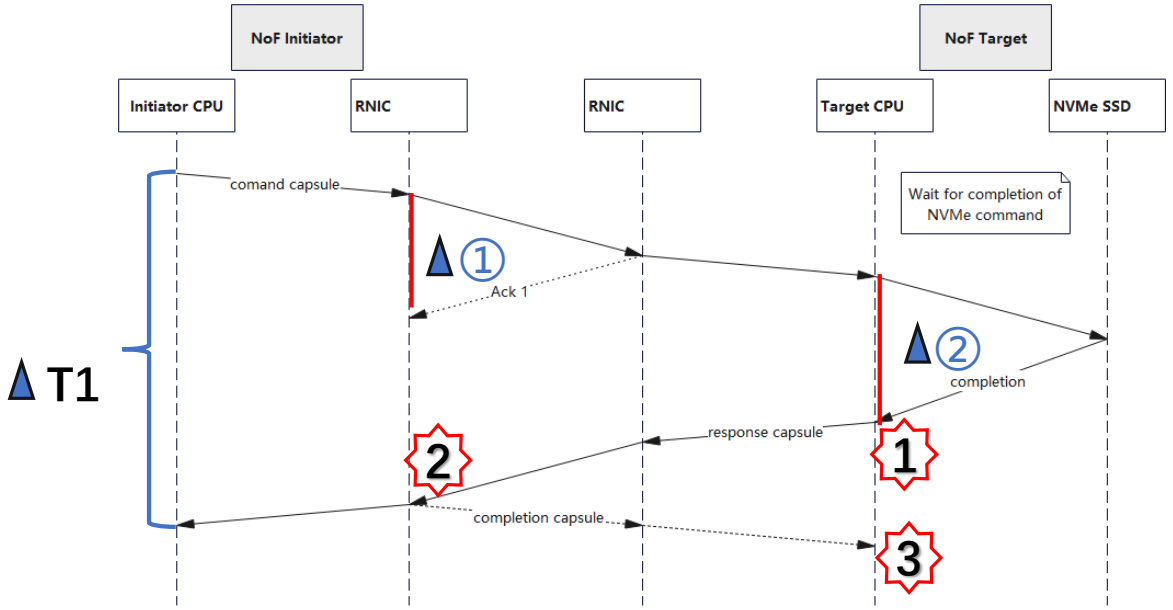
- 使用二层可靠网络优化三层应用



通过PMEM把SSD持久域扩展到Target整板存储空间

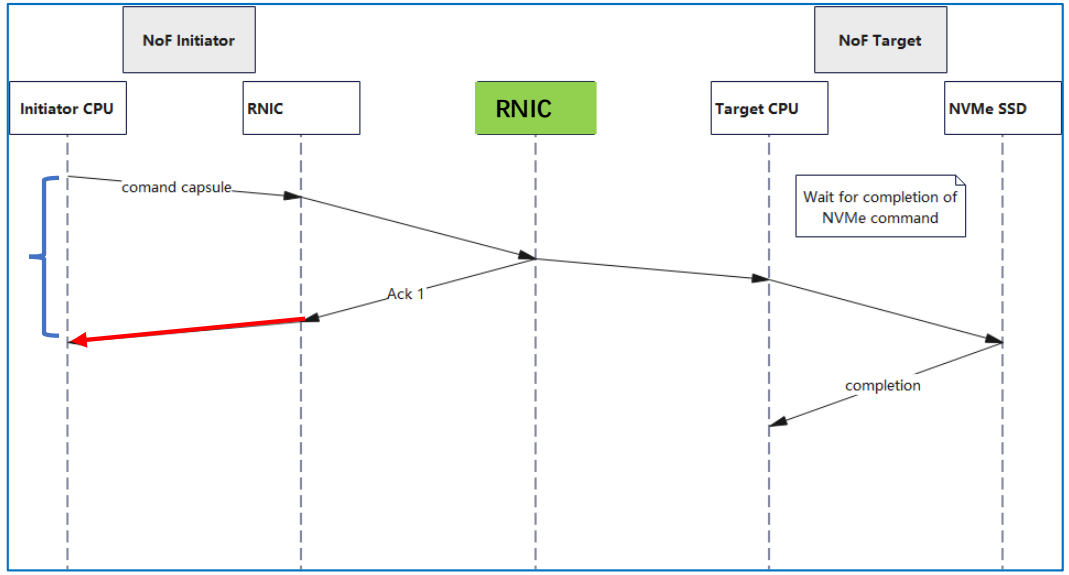
2 aNOF方案 —— 协议优化实现极致低时延

基线方案：



T1

优化方案：



T2

- **零 NVMe落盘时延**
- **零 CPU响应时延**
- **网络交互2次 → 1次**

注：

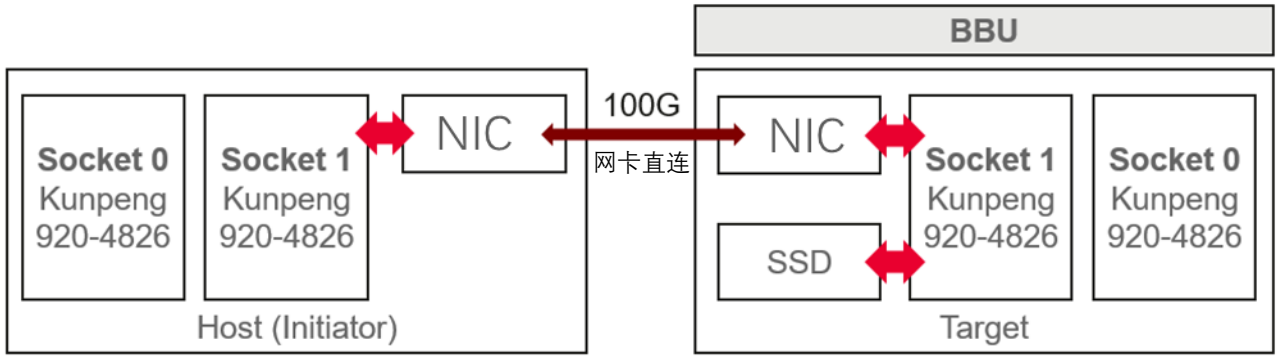
①：一次网络交互时延

②：NVMe SSD落盘时延

1 2 3：网络交互CPU开销

3 DEMO验证结果

验证环境配置：

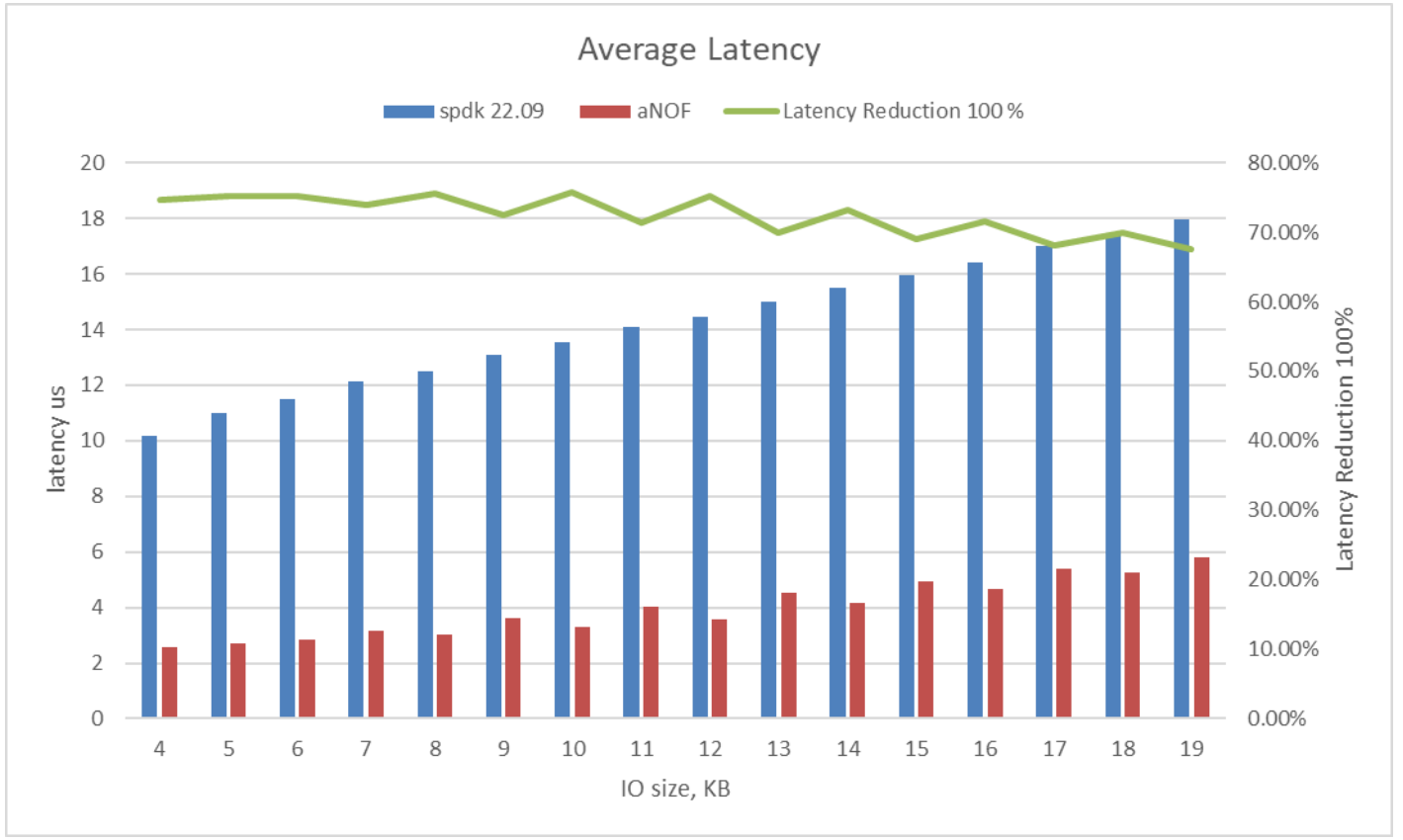


Test setup scheme

Test setup configuration

Network	
NICs	kunpeng Integrated NIC
Operation mode	RoCE
Host (initiator)	
CPU model	Kunpeng 920-4826
Number of sockets	2
SMMU	On
Prefetchers	Off
OS	openEuler 22.09
spdk	22.09
Target	
CPU model	Kunpeng 920-4826
Number of sockets	2
SMMU	On
Prefetchers	Off
SSD	3x ES3000 V6 NVMe SSD
PMEM	BBU PMEM (64G)
OS	openEuler 22.09
spdk	22.09

3 DEMO验证结果



- 1，aNOF 4k小块写瓶颈时延低于3us（实验室数据，网卡直连）
- 2，aNOF相比spdk22.09的小块写时延平均降低70%


OpenEuler